IDENTIFICATION OF MORTGAGE DEMAND FUNCTION WITH HETEROGENEOUS PREFERENCES[1]

Evgeniy M. Ozhegov[2]

National Research University - Higher School of Economics

This paper analyzes the mortgage borrowing process from a Russian state-owned provider of residential housing mortgages concentrating on the estimation of demand function with heterogeneous borrowers' preferences. Analysis takes into account the underwriting process and the choice of contract terms of all loans originated from 2008 to 2012. Our dataset contains demographic and financial characteristics for all applications, loan terms and the performance information for all issued loans by one regional bank that operates government mortgage programs.

We use a multistep semiparametric approach to estimate the determinants of bank and borrower choice controlling for possible sample selection and endogeneity of contract terms. The main contribution to the literature is modeling choice of contract terms as interdependent by structural system of simultaneous equations with heterogeneous marginal effects.

We found that demand of low-income households who are unable to afford improving of housing conditions by other instruments than government mortgage is less elastic according both to the change in interest rate and maturity.

JEL classification: C14, C30, C51, G21.

Keyword: mortgage, demand, terms choice, semiparametrics.

[1] This study (research grant No 14-01-0104) was supported by The National Research University–Higher School of Economics' Academic Fund Program in 2014- 2015

[2] Department of Economics, Research group for applied markets and enterprises studies, Perm, 614022, Lebedeva st., 27. E-mail: tos600@gmail.com

1. Introduction

During the previous decade, the Russian housing market was affected by two events. First, the worldwide financial crisis caused housing prices to fall by 30% in September 2009 compared with prices in July 2008[3]. The volume of loans issued in 2009 fell to 25% of the level of 2008. Secondly, Agency of Housing Mortgage Lending (AHML), the state-owned mortgage provider, increased the volume of mortgages issued by 120% from 2006 to 2012 without any spillover during the crisis[4] and now holds 7-12% of the market share. This means that demand for government-issued loans is rising despite the rises and falls in the economy and financial markets.

The government mortgage lending system in Russia is based on the principles of two-level system. National institute for development of housing activity - Agency of Home Mortgage Lending (AHML) was set up in 1997 as an analog of Fannie Mae and the first steps were taken towards the introduction of mortgage securities in Russia. AHML is a government sponsored enterprise (GSE) (Khmelnitskaya, 2014). It helps to implement strong government housing policy and anti-recessionary measures to support mortgage lending in Russia. In essence, AHML is the national regulator of the mortgage market.

AHML uses two-level system of lending. In the first step banks and non-credit organizations provide mortgage loans to households according the common standards of AHML. The second step is refinancing (redemption) of mortgage receivables by AHML. AHML develops both nonspecial and special mortgage programs and refinances risks from its regional branches and commercial banks, which operates such programs. The list of special programs contains "Young researchers", "Young teachers", "Mortgage for Soldiers", "Mothers' capital" and other social and subprime programs which are linked with low- and middle-income households in Russia who are usually unable to afford improvement of housing conditions using commercial mortgage instruments or savings. This makes a sample of AHML heterogeneous by socio-demographic characteristics of borrowers and, possibly, by their preferences on loan terms. Low-income borrowers usually have lower downpayment, which leads to increase of interest rate and compensation of higher annual income with longer maturities. This may make demand for mortgage loans of low-income households less elastic by interest rate comparing with higher-income households.

We are interested in the estimation of demand function for government mortgages in Russia focusing on the heterogeneity of borrowers preferences. This paper uses unique loan-level data on

[3] By the Indicators of housing market's Price Index, www.irn.ru
[4] Agency of Housing Mortgage Lending data, www.ahml.ru

applications and originated loans from one regional AHML subsidiary. Along with the heterogeneity of preferences, we also control for the selection bias that arises during the underwriting process and simultaneity in the choice of all loan terms.

Next section describes borrowing process in AHML and modeling issues. Section 3 details the data. Section 4 contains the econometric model. Section 5 describes the results of estimation. Section 6 concludes.

2. Theoretical background

The demand for mortgages from a particular bank is usually considered as a dependence between loan size and characteristics of a potential borrower and credit terms (Ozhegov, 2014).

This research analyzes the demand for mortgages from a Russian regional bank which offers mortgage programs developed by AHML. AHML is a fully state-owned company which develops nonspecial mortgage programs and programs for special groups of borrowers ("young families", "young teachers", "soldiers" and so on) and higher risk borrowers who are unable to get a mortgage from commercial banks. These programs are developed for commercial banks and AHML regional agents who originate AHML loans. If a bank issues a mortgage on the AHML program with documentation satisfying the "AHML Standards", then this loan will be automatically refinanced by AHML. The bank is paid a fixed reward for originating a loan and annual payments for operating relationships with borrowers.

The borrowing process from a point of borrower has 3 steps:

1. Application

A potential borrower chooses a credit organization and credit program that reflects their preferences, fills out an application form with their demographic and financial characteristics.

2. Credit underwriting

Considering the application and recent credit history, the credit organization approves or disapproves the application, inquires the form data.

3. Contract agreement and choice of credit terms

The approved borrower chooses a particular property to buy and credit terms: loan amount, downpayment, monthly payment and maturity. The interest rate is determined by the credit program and depends on the other terms.

Traditional models for demand estimation on the residential mortgage market used a parametric approach to estimate the contract terms choice, usually loan amount or LTV. The two

main challenges for those models have been widely discussed. The first is sample selection and the second is the endogeneity of the other contract terms.

Sample selection issues arise when decisions on a loan are made sequentially and some explanatory variables are partially observed at various stages of the lending process. If the approval process is correlated with the choice of contract terms then the magnitude of bias depends on the strength of the correlation between the LTV choice and the underwriting process, and also on the available data in the application sample (Ross 2000).

Mortgage borrowing as a sequence of consumer and bank decisions was introduced by Follain (1990). He defines the borrowing process as the choice of how much to borrow, if and when to refinance or default, and the choice of mortgage instrument itself. Rachlis and Yezer (1993) suggest a theoretical model of the mortgage lending process, which consists of a system of four simultaneous equations: (1) borrower application, (2) borrower selection of mortgage terms, (3) lender endorsement, and (4) borrower default. This paper investigates the nature of the inconsistency of estimates of recent research on borrower discrimination and shows that all four equations (and decisions) should be considered as interdependent.

Public data, such as American mortgage datasets from the Federal Housing Authority (FHA) foreclosure, The Boston Fed Study, The Home Mortgage Disclosure Act (HMDA) was published in the middle of 1990s. Using this data a few empirical studies analyzed the mortgage lending process and studied the interdependency of bank endorsement decisions and borrower decisions modeled by the bivariate probit model. As an extension of the study Rachlis and Yezer, (1993), Yezer, Philips and Trost (1994) applied a Monte-Carlo experiment to estimate the theoretical model. They empirically show that isolated modeling of the processes of credit underwriting and default lead to biased parameter estimates. Phillips and Yezer (1996) and Munnell, Tootell, Browne and McEneaney (1996) supported these findings.

Later papers studied the dependence of credit terms choices on the other endogenous variables. Ambrose, LaCour-Little and Sanders (2004) outlined the endogeneity of the loan amount and LTV.

As key determinants for the demand for the residential mortgages, authors usually select socio-demographic characteristics of borrower and contract terms. Bajari, Chu and Park (2008) also use district-level aggregated demographic and economic variables as proxies for individual characteristics when they are unavailable.

Attanazio, Goldberg, Kyriazidou, (2008) applied three-step nonparametric approach to estimate the demand for car loans corrected for sample selection and the endogeneity of rate and maturity. First, they estimated the probability of taking a loan, then reduced form residuals from the interest rate and maturity equations and then the demand equation corrected for sample

selection and endogeneity of contract terms. They found empirical evidence of nonlinearity in the demand function and the non-normality of the joint distribution of error terms.

Nonparametric regressions allow to use more flexible functional form of regression functions and control for heterogeneity of borrowers. (Attanasio et al., 2008) shows that rate and maturity elasticities of LTV vary over the income distribution of borrowers. DeFusco and Paciorek (DeFusco, Paciorek, 2014) who studied interest rate elasticity of mortgage loan amount supports this result. They showed that splines approximation helps to control for nonlinear dependency of LTV and interest rate, especially for the discontinuity of regression function near conforming loan amount limit.

However, not only loan amount depends on the other contract terms. The choice of loan amount and LTV may also affect all the other contract terms. Higher risk loans relate to the credit programs with a higher rate. A higher loan amount with a fixed rate requires larger monthly payments for unconstrained borrowers and maturity extension for credit constrained borrowers (Attanasio et al., 2008). That is the reason why choice of all credit terms in this research is modeled as structurally interdependent and we should account for the heterogeneous preferences of mortgage terms.

Recent papers showed that the approval process affects borrower decisions. Table 1 also gives evidence of a biased sample of the characteristics of the borrowers who did not sign a credit contract. In general, when modeling the contract term choice we may consider the subsample as nonrandom and biased because: 1) some applicants were considered uncreditworthy and rejected; 2) some approved borrowers did not sign a contract because of better alternatives in other banks. With the data available, we cannot separate these two reasons since we do not know the approval decision for all the applicants who did not sign a contract.

To sum up, borrowing process is represented by the following econometric model:

$$\begin{aligned} d_i &= \begin{cases} 1, g_0(x_i, z_{0i}) + e_{0i} \geq 0 \\ 0, g_0(x_i, z_{0i}) + e_{0i} < 0 \end{cases} \\ &\begin{cases} y_{1i}^* = g_1(y_{-1i}^*, x_i, z_{1i}) + e_{1i} \\ \dots \\ y_{ki}^* = g_k(y_{-ki}^*, x_i, z_{ki}) + e_{ki} \end{cases} \\ y_i &= \begin{cases} y_i^*, \text{if } d_i = 1 \\ \text{is unobserved, otherwise} \end{cases} \end{aligned} \tag{1}$$

where d_i is a binary indicator of taking a loan, x_i is a set of demographic and financial characteristics of the borrower and co-borrowers, y_i is the set of credit terms including logarithm of loan amount, LTV, logarithm of interest rate and logarithm of maturity, z_{0i} is an excluded

variables for the taking a loan decision and $z_i = (z_{1i}, \dots, z_{ki})$ is the set of excluded instruments for credit terms.

3. Econometric model

3.1. Identification

Model (1) contains a system of simultaneous equations when we model the choice of contract terms. Moreover, contract terms are observable only for the subsample of borrowers who have signed contract. This means that we possibly have selection bias problem.

The sample selection bias problem initially was discussed in Gronau (1973) and Heckman (1974). Heckman proposed methods to estimate these models using maximum likelihood or the two-step procedure in Heckman (1976, 1979) which corrects the error term in the outcome equation on covariance with the selection equation error term. However, both approaches have been limited by an assumption on the joint error distribution. Further papers deal with a relaxation of the distribution assumption for the two-step procedure using a non(semi-)parametric approach for model estimation, for instance, using a Fourier decomposition of unknowns in terms of a functional form error correction function (Heckman and Robb, 1985), or an approximation by a series of power functions (Newey, 1988).

While modeling the borrower choice of contract terms we need to allow regressors to be endogenous and represent the system of equations for each endogenous variable in structural form. Economics theory not restricts regression functions by any assumptions. Newey and Powell (1989) introduced a nonparametric procedure for the estimation of a triangular system of simultaneous equations with unknown regression functions. Newey, Powell and Vella (1999) proposed a two-step procedure for the correction of an error term on the endogeneity of regressors approximating the control function by power series on reduced form residuals. Then Newey (2013) provided an overview of nonparametric instrumental variable methods for simultaneous equations and discussed the problem of weak instruments.

Das, Newey and Vella (2003) proposed a model with both sample selection and endogenous regressors for the case of one equation of interest, and its estimation procedure. They also approximated a control function using a power series which depended on the propensity score from the selection equation and the endogenous variable reduced form equation residuals.

We extend the proposed methods for the consistent estimation of a non-triangular system of simultaneous equations with sample selection, endogenous regressors and arbitrary joint error

distribution and the functional form of regression and the control functions in reduced and structural forms. We may apply this method to estimate model (1) with the following steps.

1. We need to estimate the propensity score for the contract agreement equation:

$$p = E[d|x, z_0] = g_0(x, z_0) \quad (2)$$

2. We estimate each contract term equation in the reduced form corrected for sample selection using estimates of propensity score::

$$E\left[y_j \middle| x, z, z_0, d = 1\right] = \gamma_j(x, z) + \lambda_j(\hat{p}) \quad (3)$$

3. We estimate the structural form contract term equations corrected for sample selection, endogeneity and simultaneity using the estimates of propensity score reduced form contract term residuals:

$$E\left[y_j \middle| y_{-j}, x, z, z_0, d = 1\right] = g_j\left(y_{-j}, x, z_j\right) + \varphi_j(\hat{p}, \hat{e}_{-j}) \quad (4)$$

Identification conditions for equations (2-4) are formulated with the following theorem.

Theorem 1. *If functions* $g_0(x, z_0)$, $\gamma_j(x, z_1, \dots, z_k)$, $g_j\left(y_{-j}, x, z_j\right)$, $\lambda_j(\hat{p})$, $\varphi_j(\hat{p}, \hat{e}_{-j})$ *are continuously differentiable with continuous distribution functions almost everywhere and with probability one* $\frac{\partial g_0(x,z_0)}{\partial z_0} \neq 0$ *and* $rank\left[\frac{\partial \gamma(x,z)}{\partial z}\right] = dim(y)$ *then each regression function in (2-4) is identified up to an additive constant.*

Proof. See appendix.

To sum up all the necessary identification conditions, the assumptions of the model restricts the regression function and control function at each step to be functions from different variables and to be separable. The control function also must be a function from the variables which were obtained from the previous steps of estimation procedure.

The first group of Theorem 1 conditions ($\frac{\partial g_0(x,z_0)}{\partial z_0} \neq 0$ and $rank\left[\frac{\partial \gamma(x,z)}{\partial z}\right] = dim(y)$) restricts the data. Thus, there must be at least one significant variable in the selection equation excluded from the system and at least one relevant excluded instrument for each endogenous variable y.

The last group restricts all regression and control functions to be continuously differentiable.

An estimation procedure is based on an approximation by a series of power functions which depend on the initial set of regressors. This family of regression functions satisfies the differentiability conditions of Theorem 1.

Let $\omega = (\omega_1, \dots, \omega_\chi)$ be a set of variables with $\chi = dim(\omega)$.

$\kappa(\rho,\chi)=\frac{(p+\chi)!}{p!\chi!}$ will be the number of polynomial terms with a power no more than ρ which may be obtained from χ variables.

Let $\mathrm{Q}^{\rho}(\omega)=(q_1(\omega),\dots,q_\kappa(\omega))$ be a vector of κ power functions, which are a full set of polynomial terms with a power no more than ρ obtained from ω, i.e. $q_j(\omega)=\prod_{\tau=1}^{\chi}\omega_\tau^{s_\tau}, \sum_{\tau=1}^{\chi}s_\tau\leq\rho$, $s_\tau\in\{0,1,\dots,p\}\ \forall\tau=\overline{1,\chi}$.

Let $\mathrm{Q}^{\rho}(\omega)$ be a polynomial approximating series with power ρ.

Consider $Q_0=Q^{\rho_0}(x,z_0)$ as selection equation approximating function.

Then the propensity score of the selection equation may be estimated by OLS as

$$\widehat{p_i}=E[d_i|x,z_0]=Q_0'[Q_0'Q_0]^{-1}Q_0'd_i \tag{5}$$

Let $Q_r=(Q^{\rho_1}(x,z),Q^{\rho_1}(\hat{p}))$ be the reduced form regression and control functions' approximating functions and $b_j=(b_{1j},b_{2j})$ be vectors of their parameters. Then b_j may be obtained by OLS as

$$\hat{b}_j=[Q_r'Q_r]^{-1}Q_r'y_j \tag{6}$$

Then the reduced form contract terms residuals will be

$$\hat{e}_{ji}=y_{ji}-Q_r'\hat{b}_j \tag{7}$$

Let $Q_j=\left(Q^{\rho_1}(y_{-j},x,z_j),Q^{\rho_1}(\hat{p},\hat{e}_{-j})\right)$ be the structural form regression and control functions' approximating functions and $\beta_j=(\beta_{1j},\beta_{2j})$ be vectors of their parameters. Then the estimate for β_j may be obtained by OLS as

$$\hat{\beta}_j=[Q_j'Q_j]^{-1}Q_j'y_j \tag{8}$$

The next theorem introduces conditions for the consistency of the proposed estimation procedure.

Theorem 2. *If equations (2-4) are identified through theorem 1 and the set of variables* (x,z,z_0) *is independent from the distribution of* $(e_0,e_1,\dots,e_k)$ *then* $\hat{b}_j$ *and* $\hat{\beta}_j$ *are consistent.*

Proof. See in Appendix.

3.2. Testing exclusion restrictions

Crucial assumption for identification is full rank of matrix of marginal effects of excluded instruments. Instead of validity assumption this one is testable. We will follow Sanderson and Windmeijer (2014) conditional *F*-test approach in order to test the hypothesis of matrix of marginal effects reduction to one from full rank.

Firstly, consider testing of linear model with multiple endogenous variables. Then we generalize the test for nonlinear semiparametric model with sample selection.

In simplest case studied by Sanderson and Windmeijer (2014) we have one linear equation of *y* with *k* endogenous variables $x = (x_1, \dots, x_k) = (x_j, x_{-j})$ and *m* instruments $z = (z_1, \dots, z_m)$ independent on distribution of error terms. The model may be expressed as

$$y = x\beta + e_0$$
$$x_j = z\gamma_j + e_j$$
$$(e_0, e_1, \dots, e_k) \perp z$$

The problem is to test whether $rank[\Gamma' z' z \Gamma] = k$. Stock and Yogo (2005) introduced a test based on minimal eigenvalue of matrix $\frac{\hat{\Gamma}' z' z \hat{\Gamma}}{m}$ with $H_0: rank[\Gamma' z' z \Gamma] = k - 1$. If its minimal eigenvalue statistically differs from zero we can reject the null that matrix has not full rank and instruments are weak. They also calculated critical values for the test but for $k \leq 2$. Sanderson and Windmeijer (2014) followed Angrist and Pischke (2009) conditional *F*-test approach for testing of joint significance of instruments in reduced form regression. Angrist and Pischke (2009) proposed conditional *F*-statistics and Sanderson and Windmeijer (2014) corrected its asymptotic distribution and proved equivalence to Stock and Yogo (2005) test. For this application conditional *F*-testing approach has an advantage of existence of known limiting distribution that can be easily extended for semiparametric and sample selection case and we can easily calculate its critical values even for $k > 2$. Testing contains 3 steps: 1) estimation of endogenous variable conditional on all other endogenous variables, 2) estimation of conditional reduced form parameters and 3) calculating test statistics. Formally saying:

1) Obtain $\hat{\kappa}_j$ by OLS from regression $x_j = x_{-j}\hat{\kappa}_j + \xi_j$
2) Obtain $\hat{\gamma}_j$ by OLS from regression $x_j - x_{-j}\hat{\kappa}_j = z\hat{\gamma}_j + \nu_j$
3) For each endogenous variable calculate instrument's conditional $F_{x_j|x_{-j}} = \frac{\hat{\gamma}_j' z' z \hat{\gamma}_j}{(m-k+1)(\frac{\hat{\nu}_j' \hat{\nu}_j}{n})}$.

For the case of semiparametric equation with continuously differentiable regression functions like

$$y = g(z) + e_0$$
$$x_j = \gamma_j(z) + e_j$$
$$(e_0, e_1, \dots, e_k) \perp z$$

for joint instrument's relevance we need to prove that $rank\left[\frac{\partial \gamma(z)}{\partial z}\right] = dim(x) = k$. If we approximate each unknown regression function $f(s)$ with its polynomial approximation function $Q^\rho(z)\alpha$ with power ρ and $dim(\alpha) = \frac{(\rho + dim(z))!}{\rho!\, dim(z)!}$ then we can test exclusion restriction by:

1) Obtain $\hat{\kappa}_j$ by OLS from regression $x_j = Q^\rho(x_{-j})\hat{\kappa}_j + \xi_j$
2) Obtain $\hat{\gamma}_j$ by OLS from regression $x_j - Q^\rho(x_{-j})\hat{\kappa}_j = Q^\rho(z)\gamma_j + \nu_j$
3) For each endogenous variable calculate instrument's conditional $F_{x_j|x_{-j}} = \frac{\hat{\gamma}_j'(Q^\rho(z))'Q^\rho(z)\hat{\gamma}_j}{(\frac{(\rho+\mathrm{m})!}{\rho!\mathrm{m}!} - \frac{(\rho+\mathrm{k})!}{\rho!\mathrm{k}!}+1)(\frac{\hat{\nu}_j'\hat{\nu}_j}{n})}$.

This type of testing is very simply may be generalized for the case of presence of sample selection by including control functions for error terms when estimating regressions parameters on steps (1-2).

4. Data description

4.1. Data set

One of the regional AHML operators provided the data set of all applications for mortgage collected from 2008 to 2012. We know the demographic and financial characteristics of each of the 3870 applicants as main borrowers and their co-borrowers on the date of application. We also know the date of application. For all signed contracts we know the loan limit set by the bank, the contract terms, and the value of property. The characteristics (Table 1) of the borrower are fully observable and the contract characteristics are partially observable for only the subsample of applicants who signed the contract.

Some mortgage programs allow the applicants not to provide any information on their income or to provide it in the form that bank refuses to approve. In this case, we does not observe income in the data. These programs are usually linked with a higher contract rate. The reason for this choice may be explained by a temporary or changeable income (LaCour-Little, 2007), for instance, for entrepreneurs. Generally, income should be considered endogenous while modeling the approval of borrower or contract terms choice. However, we can control for employment category, which rejects the inconsistency due to possible endogeneity of income.

Moreover, co-borrower income may also be endogenous and we cannot provide any proxy for co-borrower income since we do not have any characteristics of co-borrowers. This is a limitation of the research. But we may consider it as insignificant for the choice of contract terms compared to the income of the main borrower.

Originated loans have high variation in all chosen terms (Table 2). AHML provides only fixed-rate mortgages in roubles. All values was converted from roubles to US dollars by the official exchange rate in the month of application.

Tab. 1. Descriptive statistics for applicants' characteristics.

Variable	All applicants (3344[5] obs.)	Did signed contract (2019 obs.)	Did not signed contract (1325 obs.)
Age[6], years	33.77 (7.56)	33.93 (7.65)	33.53 (7.41)
Sex			
Male	1848 (55.3%)	1151 (57.0%)	697 (52.6%)
Female	1496 (44.7%)	868 (43.0%)	628 (47.4%)
Marital status			
Married	1793 (53.6%)	1132 (56.1%)	661 (49.9%)
Single	1013 (30.3%)	587 (29.1%)	426 (32.2%)
Divorced	497 (14.9%)	281 (13.9%)	216 (16.3%)
Widowed	41 (1.2%)	19 (0.9%)	22 (1.7%)
Category of employment			
Hired employee	3210 (96.0%)	1923 (95.2%)	1287 (97.1%)
State-owned employee	111 (3.3%)	79 (3.9%)	32 (2.4%)
Entrepreneur	23 (0.7%)	17 (0.8%)	6 (0.5%)
Level of education			
Complete higher	1756 (52.5%)	1116 (55.3%)	640 (48.3%)
Less than higher	1588 (47.5%)	903 (44.7%)	685 (51.7%)
Declared income of main borrower			
Not declared	2333 (69.8%)	1223 (60.6%)	1110 (83.8%)
From 0 to $249	85 (2.5%)	47 (2.3%)	38 (2.9%)
From $250 to $499	279 (8.3%)	237 (11.7%)	42 (3.2%)
From $500 to $1 000	442 (13.2%)	358 (17.7%)	84 (6.3%)
More than $1 000	205 (6.1%)	154 (7.6%)	51 (3.8%)
Number of co-borrowers			
0	1416 (42.3%)	823 (40.8%)	593 (44.8%)
1	1794 (53.6%)	1105 (54.7%)	689 (52.0%)
2	134 (4.0%)	91 (4.5%)	43 (3.2%)
Declared income of co-borrowers			
Not declared	2939 (87.9%)	1677 (83.1%)	1262 (95.2%)
From 0 to $249	105 (3.1%)	97 (4.8%)	8 (0.6%)
From $250 to $499	157 (4.7%)	129 (6.4%)	28 (2.1%)
More than $500	143 (4.3%)	116 (5.7%)	27 (2.0%)

Tab. 2. Descriptive statistics of the issued loans (2019 contracts).

Variable	Mean	St. dev.	Min	Max
Loan amount, $	25 068.3	12 440.54	3 750	120 000
Downpayment, $	20 130.0	13 740.91	1 250	117 500
Flat value, $	45 198.3	21 191.38	12 500	175 000
Monthly payment, $	303.5	158.8	60.2	1 766
Loan-to-value ratio (LTV), %	57.2	16.1	11.0	94.3
Maturity, months	190.4	61.5	26	360
Rate, %	11.4	1.54	9.5	19

[5] The outliers from the sample were excluded. We treat an observation as an outlier if the age, level of education, marital status or other characteristics were missing (119 obs.). We exclude observations with borrowers under age 21, with LTV or DTI (debt-to-income ratio) less than 0 or more than 1 (135 obs.). We consider those outliers as random and due to the errors in the database. We also exclude 2.5% observations with extremal value of bought property from each side of its distribution (89 obs.). After excluding the outliers the sample was 3344 observations. 2019 applicants signed the mortgage contract, while 1325 of them did not.

[6] Mean and standard deviation in the parenthesis.

4.2. Instrumental variables

To estimate the model we need to find a set of relevant excluded instruments for the probability of signing a contract and each credit term.

Bajari et al. (2008) discussed the possibility of using aggregated district-level variables as proxies for unavailable data. We will use the same strategy to find the set of instruments. Since we have data without spatial variation we can use time variation in applications. We have data on applications from July 2008 to August 2012 and we know the application date for each applicant. Each application was matched with the set of aggregated mortgage and housing market characteristics for the same month. On average, the process takes two months from the date of application to the date of contract agreement. Also Ozhegov and Poroshina (2013) showed that aggregated demand on mortgage reacts to changes in supply within two months. Then we need to fix the aggregated market characteristics for each application not only in the month of application, but also the 1-2 months prior the application, and use these as instruments.

Table 3 represents the descriptive statistics of aggregated mortgage and housing market characteristics for the period from July 2008 to August 2012 (50 months).

Tab. 3. Aggregated mortgage and housing markets characteristics.

Variable	Mean	St. dev.	Min	Max
Volume of issued mortgage in region, mln. $	23.0	14.1	2.9	54.8
Volume of issued mortgage in region, number	894.4	529.2	134	2112
Mean loan amount, $	28 814.2	6299.8	22 482.7	47 705.0
Median maturity, months	200.79	12.81	173	222.2
Median rate, %	12.97	0.80	12	14.3
Mean LTV	0.58	0.03	0.48	0.65
Mean DTI[7]	0.35	0.01	0.33	0.37
Mean ft^2 value, $	89.7	14.3	66.9	119.2
Affordability of housing coefficient[8]	0.287	0.055	0.215	0.389
Number of refinanced in AHML loans	129.1	83.7	30	310
Number of application to the bank	121.4	51.9	43	222

About 15% of issued loans were refinanced by AHML, but not all of them were issued by the bank supplying the data. Generally, the number of applications to the bank is fewer than the number refinanced by AHML by all the regional banks.

The difference between the number of loans refinanced by AHML and the number of applications to the bank within a particular month may be the excluded variable which explains the probability of contract agreement, but it does not affect the contract term choice. Since every commercial bank operates with the same AHML programs, the difference in the approval process

[7] DTI – ratio between monthly payment and monthly income.
[8] Affordability coefficient reflects the ratio between an annual income of mean household and a value of mean flat.

does not affect the terms choice. But an increase in the number of refinanced loans shows the changes in the underwriting process in other banks and may correlate with the probability of a contract agreement with the bank. This variable should be considered as exogenous since each individual decision explains a negligible variation of the aggregated market characteristic (less than 1%).

As excluded instruments for credit terms, loan amount, LTV, interest rate, maturity, we used mean LTV and mean DTI (Debt-to-Income ratio), median rate, median maturity for originated loans in region and the housing affordability coefficient. The relevance of the instruments set is implied by the interdependence of mortgage market characteristics and the AHML credit programs conditions. Validity is implied by the exogeneity of the program terms in respect to each particular borrower. The relevance was proved for each model with the *F*-test for the excluded instrument in section 5.

5. Results

Model (1) was estimated with the proposed procedure (5-8).

First, we estimated the model of the probability of a contract agreement (Tab.A.1.) based on the characteristics of the borrower and co-borrowers and the difference between the number of AHML refinanced loans and the number of applications. The last variable which was taken as an excluded instrument is significant at the 1% level. The sign and significance of borrower characteristics are consistent with recent research. The demographic characteristics, such as age, sex, marital status of the borrower are insignificant, which supports the absence of discrimination. However, borrowers are discriminated by level of education. The probability of a contract agreement is positively correlated with the income of the main borrower and, on the contrary, negatively correlates with the failure to provide income details. Entrepreneurs have a higher probability of a contract agreement *ceteris paribus*.

These estimates were obtained from the linear probability model and were compared with the probit model. The comparison showed an insignificant difference in the significance of the parameter estimates and predictive power (with slightly higher predictive power for the linear probability model). The propensity score $\widehat{p_i} = E[d_i|x_i, z_{0i}]$ was obtained from the linear probability model.

For each credit term we estimated the reduced form equation. The control function was approximated by the polynomial with power ρ_1 on the estimate of the propensity score. The

regression function was estimated as partially polynomial. It was linear for the characteristics of the borrower and polynomial for the excluded instruments for contract terms with power ρ_1. We test three set of instruments described earlier. First, we fix market-level variables on the month of application. For the second and third sets, we used market-level data for month one and two months fore the month of application respectively. The proof of relevance of excluded instruments based on conditional *F*-test is provided in Table 4.

All sets of excluded instruments are relevant on 1% level. We use then the reduced form residuals obtained from models (II) because of the best joint approximation of endogenous variables variance.

Tab. 4. Results of instruments' relevance test

	(I)			(II)			(III)		
Equation	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Log. of loan limit	7.57	5.06	3.41	5.66	5.08	3.30	7.59	4.44	3.02
LTV	2.19	2.35	2.25	4.18	2.38	2.22	2.89	2.29	2.53
Log. of rate	143.6	65.2	22.2	156.9	64.8	28.7	174.8	73.0	35.1
Log. of maturity	5.61	2.42	2.01	7.13	2.94	2.09	6.48	2.65	2.08
10% critical values	1.40	1.34	1.25	1.40	1.34	1.25	1.40	1.34	1.25
5% critical values	1.55	1.46	1.33	1.55	1.46	1.33	1.55	1.46	1.33
1% critical values	1.84	1.69	1.49	1.84	1.69	1.49	1.84	1.69	1.49

Note: In the table cells there are conditional *F*-statistics of excluded instruments.
Critical values are provided.
For each equation, models (I) are calculated with market-level instruments fixed in the month of application, models (II) with market-level instruments fixed one month before the month of application, and models (III) for two months before the month of application.
For each equation, model (1) was estimated for $\rho_1 = 1$, model (2) for $\rho_1 = 2$, model (3) for $\rho_1 = 3$.

We estimated the contract term equations in structural form using a polynomial approximation with power ρ_1 for the control function on $\hat{p}$ and the reduced form of the contract term equations. The regression function was partially polynomial, linear for the characteristics of the borrower and polynomial with power ρ_1 for the credit terms. First we was focused on the necessity of the estimates correction for simultaneity in choice of loan terms and endogenous selection of borrowers. Table 5 presents results for loan amount equation estimates using models with different approximation power. Each model was estimated with and without control for simultaneity and sample selection. Following Hausman test approach we test the difference of estimates with and without control which shows inconsistency of estimates without control if the difference is significant (p_1 for H_0: difference in estimates is not systematic).

Tab. 5. Comparison of full model with models without correction

	(I)				(II)				(III)			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
LTV	0.020***	0.020***	0.010***	0.010***	0.020***	0.021***	0.011***	0.010***	0.020***	0.020***	0.010***	0.009***
	(0.002)	(0.002)	(0.002)	(0.001)	(0.004)	(0.003)	(0.003)	(0.003)	(0.005)	(0.04)	(0.04)	(0.04)
Log. of rate	-0.575***	-0.497***	-0.438***	-0.402***	-0.567***	-0.492***	-0.352***	-0.346***	-0.599***	-0.536***	-0.471***	-0.499***
	(0.047)	(0.059)	(0.060)	(0.040)	(0.080)	(0.070)	(0.065)	(0.063)	(0.149)	(0.072)	(0.090)	(0.099)
Log. of maturity	0.479***	0.369***	0.285***	0.286***	0.471***	0.357***	0.271***	0.274***	0.487***	0.357***	0.263***	0.268***
	(0.058)	(0.087)	(0.061)	(0.040)	(0.090)	(0.038)	(0.044)	(0.041)	(0.147)	(0.112)	(0.062)	(0.060)
Age of borrower	0.020***	0.013***	0.018*	0.018*	0.022***	0.013***	0.019*	0.0194*	0.020***	0.013***	0.020*	0.021*
	(0.002)	(0.003)	(0.008)	(0.008)	(0.003)	(0.003)	(0.008)	(0.008)	(0.003)	(0.003)	(0.008)	(0.008)
Age squared	-0.000**	-0.000	-0.000	-0.000	-0.000**	-0.000	-0.000	-0.000	-0.000*	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Male	-0.036***	-0.028***	-0.024	-0.021	-0.036***	-0.028***	-0.024	-0.022	-0.035***	-0.028***	-0.025	-0.023
	(0.003)	(0.003)	(0.017)	(0.017)	(0.003)	(0.003)	(0.017)	(0.017)	(0.003)	(0.004)	(0.017)	(0.017)
Family status (Married is base level):												
Single	-0.052***	-0.054***	-0.036	-0.041	-0.057***	-0.056***	-0.043	-0.046*	-0.052***	-0.053***	-0.038	-0.043
	(0.004)	(0.005)	(0.023)	(0.023)	(0.005)	(0.005)	(0.023)	(0.022)	(0.004)	(0.005)	(0.023)	(0.022)
Divorced	-0.075***	-0.077***	-0.056*	-0.061*	-0.080***	-0.079***	-0.061*	-0.065*	-0.075***	-0.076***	-0.061*	-0.066*
	(0.005)	(0.006)	(0.027)	(0.027)	(0.006)	(0.006)	(0.027)	(0.027)	(0.005)	(0.006)	(0.028)	(0.027)
Widowed	-0.072***	-0.096***	-0.063	-0.083	-0.087***	-0.099***	-0.088	-0.099	-0.071***	-0.097***	-0.074	-0.095
	(0.016)	(0.017)	(0.082)	(0.081)	(0.016)	(0.017)	(0.083)	(0.080)	(0.016)	(0.017)	(0.083)	(0.080)
Category of activity (Hired employee is base level):												
Entrepreneur	0.067***	0.072***	0.066	0.066	0.067***	0.071***	0.064	0.064	0.062***	0.069***	0.077	0.078
	(0.017)	(0.017)	(0.085)	(0.085)	(0.017)	(0.018)	(0.084)	(0.085)	(0.016)	(0.017)	(0.084)	(0.084)
State employee	-0.142***	-0.091***	-0.123**	-0.094*	-0.130***	-0.089***	-0.109*	-0.091*	-0.125***	-0.089***	-0.109*	-0.088*
	(0.009)	(0.008)	(0.045)	(0.040)	(0.009)	(0.008)	(0.047)	(0.039)	(0.009)	(0.008)	(0.047)	(0.039)
Level of education (Complete higher is base level):												
Less than higher	-0.133***	-0.156***	-0.141***	-0.152***	-0.136***	-0.156***	-0.143***	-0.150***	-0.135***	-0.156***	-0.144***	-0.153***
	(0.004)	(0.003)	(0.018)	(0.016)	(0.004)	(0.003)	(0.018)	(0.016)	(0.004)	(0.003)	(0.018)	(0.016)

Tab. 5. Comparison of full model with models without correction (continuing)

	(I)				(II)				(III)			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Number of co-borrowers (No co-borrowers is base level):												
1 co-borrower	0.086***	0.095***	0.081***	0.083***	0.086***	0.093***	0.077***	0.077***	0.089***	0.096***	0.074**	0.075***
	(0.004)	(0.005)	(0.023)	(0.023)	(0.005)	(0.005)	(0.023)	(0.023)	(0.005)	(0.005)	(0.023)	(0.023)
2 co-borrowers	0.123***	0.144***	0.131**	0.133**	0.115***	0.138***	0.120**	0.120**	0.125***	0.144***	0.118**	0.118**
	(0.010)	(0.010)	(0.044)	(0.044)	(0.010)	(0.010)	(0.044)	(0.044)	(0.010)	(0.010)	(0.044)	(0.044)
Declared income of co-borrowers (Not declared is base level):												
From $0 to $249	-0.216***	-0.167***	-0.129**	-0.107**	-0.233***	-0.168***	-0.125**	-0.104*	-0.242***	-0.165***	-0.119*	-0.086*
	(0.010)	(0.010)	(0.045)	(0.041)	(0.011)	(0.010)	(0.046)	(0.041)	(0.011)	(0.010)	(0.047)	(0.041)
From $250 to $499	-0.015	0.015	0.023	0.034	-0.023**	0.014	0.028	0.037	-0.021**	0.016*	0.030	0.042
	(0.008)	(0.007)	(0.038)	(0.036)	(0.008)	(0.008)	(0.037)	(0.036)	(0.008)	(0.008)	(0.037)	(0.036)
More than $500	0.067***	0.098***	0.157***	0.167***	0.054***	0.095***	0.159***	0.168***	0.054***	0.094***	0.165***	0.174***
	(0.009)	(0.009)	(0.038)	(0.038)	(0.009)	(0.009)	(0.038)	(0.037)	(0.009)	(0.009)	(0.038)	(0.037)
Declared income of main borrower (Not declared is base level):												
From $0 to $249	-0.419***	-0.481***	-0.650***	-0.657***	-0.381***	-0.469***	-0.616***	-0.625***	-0.388***	-0.484***	-0.625***	-0.637***
	(0.017)	(0.018)	(0.056)	(0.056)	(0.019)	(0.018)	(0.057)	(0.056)	(0.018)	(0.0190)	(0.058)	(0.057)
From $250 to $499	-0.413***	-0.354***	-0.458***	-0.416***	-0.398***	-0.351***	-0.432***	-0.404***	-0.392***	-0.357***	-0.432***	-0.402***
	(0.009)	(0.008)	(0.044)	(0.029)	(0.009)	(0.008)	(0.044)	(0.029)	(0.009)	(0.00775)	(0.046)	(0.029)
From $500 to $999	-0.122***	-0.057***	-0.105**	-0.068**	-0.113***	-0.056***	-0.089*	-0.065**	-0.100***	-0.0580***	-0.089*	-0.066**
	(0.008)	(0.005)	(0.037)	(0.024)	(0.008)	(0.005)	(0.039)	(0.024)	(0.008)	(0.00502)	(0.040)	(0.024)
More than $1000	0.255***	0.316***	0.321***	0.346***	0.256***	0.315***	0.339***	0.354***	0.273***	0.320***	0.350***	0.359***
	(0.010)	(0.008)	(0.038)	(0.032)	(0.009)	(0.008)	(0.038)	(0.032)	(0.010)	(0.008)	(0.040)	(0.032)
p1	-	0.000	0.000	0.000	-	0.000	0.000	0.000	-	0.000	0.000	0.000
p2	0.000	0.000	0.032	-	0.000	0.000	0.064	-	0.000	0.000	0.141	-

Note: In the table cells there are mean marginal effects of changing of log. of loan amount. Bootstrap standard errors for 100 replications clustered on the month of application are in the parenthesis.
Significance level obtained from bootstrap distribution, * - 10%, ** - 5%, *** - 1%. 2041 observations.
Models (I) was estimated for $\rho_1 = 1$, models (II) for $\rho_1 = 2$ and models (III) for $\rho_1 = 3$.
(1) is a model controlled for both sample selection and endogeneity of contract terms, (2) is a model controlled for sample selection only,
(3) is a model controlled for endogeneity of contract terms only, (4) is a model without control for sample selection and endogeneity.
$p1$ is a p-value of Hausman test for the difference in mean marginal effects between the full model and models without control. $p2$ is a p-value of F-test of joint significance of control function parameters.

Necessity of control is also tested using Darbin-Wu approach for joint significance of control function parameters (p_2 H_0: parameters of control function are all 0).

Both tests supports the assumption that in order to obtain consistent estimates we need to control for both sample selection and simultaneity of terms choice. Otherwise, we underestimate the effect of contract terms in demand function and miss some causal relationships of borrowers' characteristics on their demand.

Estimates of demand function are staying stable with the increase of approximation power, but losing efficiency. Mean marginal effects are consistent with intuition and recent literature. Thus, demand increases with the increase of LTV and maturity and decreases with higher interest rate. Larger loans are demanded by married borrowers comparing with singles, by women, lower-aged borrowers, higher educated people. Loan amount also increases with all measures of income: higher income of main borrower and co-borrowers and with the number of co-borrowers.

Mean preferences on main mortgage charateristics such as LTV, interest rate and maturity are not counterintuitive. However, the choice of this charactersitics itself may be affected by the choice of loan amount. Estimates of structural interdependence of credit terms are showed in Table 6. Thus, LTV is higher when rate is lower and when loan amount is higher. Interest rate will increase with higher LTV and longer maturity, which is consistent with mortgage programs design. The choice of longer maturity is linked with larger loans and interest rate, which supports the recent results that maturity is very flexible instrument of monthly payments adjustments for borrowers with credit constraints (Attanasio, et. al. 2008). Moreover, from (Attanasio, et. al. 2008) we know that lower-income car credit borrowers are elastic by maturity and inelastic by rate while higher-income borrowers are inelastic by maturity and elastic by rate. This is interesting to look at the distributions of demand elasticities on interest rate and maturity along with the interdependence of this two elasticities with borrowers' income.

Graph 1 in Appendix represents the distribution of interest rate elasticity of demand for models with different approximation power. While mean rate elasticity is estimated as [-0.57, -.60], the 95% confidence interval of its distribution across borrowers sample varies 1.5 times from -0.5 to -0.75. However, we did not find any support of the presence of rate inelastic borrowers in our sample. Graph evidences the same for maturity elasticity, which varies over the sample from 0.42 to 0.56 with the mean on 0.48.

While there are no inelastic for loan characterisitcs borrowers, we may also study the interdependence of interest rate and maturity elasticity. (Attanasio, et. al. 2008) shows that there should be negative dependence in absolute values between the interest rate elasticity and maturity elasticity. Thus, less rate elastic borrowers should have higher maturity elasticity and *vice versa*.

Tab. 5. Estimates for the contract terms equations in structural form.

	Eq. 1. Log. of loan amount			Eq. 2. LTV			Eq. 3. Log. of rate			Eq.4. Log. of maturity		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Log. of loan amount		-		0.128*** (0.011)	0.125*** (0.21)	0.127*** (0.27)	-0.080* (0.052)	-0.061 (0.051)	-0.029 (0.077)	0.328*** (0.030)	0.307*** (0.081)	0.318*** (0.116)
LTV	0.020*** (0.002)	0.020*** (0.004)	0.020*** (0.005)		-		0.019*** (0.002)	0.020*** (0.003)	0.019*** (0.008)	-0.000 (0.000)	-0.000 (0.001)	-0.000 (0.001)
Log. of rate	-0.575*** (0.047)	-0.567*** (0.080)	-0.599*** (0.149)	-0.148*** (0.048)	-0.143*** (0.052)	-0.135** (0.060)		-		0.142*** (0.020)	0.153*** (0.063)	0.158** (0.073)
Log. of maturity	0.479*** (0.058)	0.471*** (0.090)	0.486*** (0.147)	-0.045* (0.031)	-0.037 (0.039)	-0.045 (0.090)	0.170*** (0.031)	0.166*** (0.032)	0.149*** (0.039)		-	
k	28	49	94	28	49	94	28	49	94	28	49	94

Note: In the table cells there are mean marginal effects of changing of dependent variable on a change of another endogenous variable.
Bootstrap standard errors for 100 replications clustered on the month of application are in the parenthesis.
Significance level obtained from bootstrap distribution,
* - 10%, ** - 5%, *** - 1%.
k – number of estimated parameters,
2041 observations.
For each equation, model (1) was estimated for $\rho_1 = 1$, model (2) for $\rho_1 = 2$, model (3) for $\rho_1 = 3$.

In order to find some evidence of this hypothesis we construct kernel regression of demand elasticities on rate and maturity for different income groups. Graph 3 in Appendix represents the estimation results. First of all, the graph rejects the assumption about negative dependence in absolute values. Over all sample of AHML borrowers two elasticities have co-movement. Thus, borrowers who are more elastic on rate are also more elastic on maturity. We also have sorting of rate elasticities across income groups with higher income linked with higher interest rate elasticity. Income level makes difference in demand elasticity on interest rate, however, this difference is statistically insignificant according to the overall elasticity confidence interval.

Russian borrowers show different behavior than those who was studied in (Attanasio, et. al., 2008). The main difference is made by the goal of taking a loan. Thus, low-income households in Russia usually can not afford either buying a new property with the savings, or buying it using commercial banks mortgage instruments. AHML special programs aim to make mortgage affordable for those households which are in need for housing conditions improvements, but can not afford it using market instruments. This makes them less elastic to the choice of characteristics of mortgage instruments and potentially pay more for the loan. But it still remains the only instrument to support housing improvements for low-income borrowers in Russia.

6. Conclusion

This paper analyzes the borrowing process in one Russian bank which is a regional subsidiary of AHML, a national provider of residential housing mortgages. This analysis takes into account the underwriting process and the choice of contract terms for all loans originated by the bank from 2008 to 2012. The dataset contains information about the demographic and financial characteristics of the borrower for all applications, the contract terms and the property value for all signed contracts. We also use regional-level aggregated housing and mortgage market characteristics as instrumental variables for the selection equation and endogenous variables and prove its relevance by conditional *F*-test.

We model the demand for loans as a simultaneous choice of loan terms and represent this as a system of simultaneous equations. We observe the choice only for those borrowers who were approved by the bank and choose to get a mortgage from this particular bank. We also consider the preferences on mortgage terms of households as heterogeneous across the sample of borrowers. This structure of borrowing process determines the use of the multistep semiparametric approach.

The main finding is that borrowers who are more elastic on interest rate are also more elastic on maturity. We also have weak evidence that lower elasticities of demand for mortgage

on interest rate and maturity are linked with low-income households who use AHML special programs as an only way to improve housing conditions.

The obtained estimates depend on the data. We used data from only one regional operator of AHML programs and do not have enough space variation. Our dataset is not big enough to apply semiparametric procedures with high-order polynomial approximations for regression and correction functions or to apply fully nonparametric model. Therefore the estimates with increasing polynomial order remains consistent but is inefficient. However, we may rely on the obtained results since the estimation procedure is based on the minimum assumptions for the consistency of estimates.

References

Andrews, D. W., Schafgans, M. M. (1998). Semiparametric Estimation of the Intercept of a Sample Selection Model. *The Review of Economic Studies, 65*(3), 497-517.

Angrist J., Pischke, J.-S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion. *Princeton University Press*, Princeton.

Ambrose, B., LaCour-Little M., Sanders, A. (2004). The Effect of Conforming Loan Status on Mortgage Yield Spreads: A Loan Level Analysis. *Real Estate Economics*, 32(4), 541–569.

Attanasio, O., Goldberg, P., Kyriazidou, E. (2008). Credit Constraints in Market for Consumer Durables: Evidence from Micro Data on Car Loans. *International Economic Review*, 49(2), 401–436.

Bajari, P., Chu, C. S., Park, M. (2008). An Empirical Model of Subprime Mortgage Default from 2000 to 2007. *NBER working paper* 14625.

Das, M., Newey, W.K., Vella, F. (2003). Nonparametric Estimation of Sample Selection Models. *The Review of Economic Studies*, 70(1), 33–58.

DeFusco, A.A., Paciorek, A. (2014). The Interest Rate Elasticity of Mortgage Demand: Evidence From Bunching at the Conforming Loan Limit. *Board of Governors of the Federal Reserve System (US)*, №2014-11.

Follain, J. R. (1990). Mortgage Choice. *Real Estate Economics*, 18(2), 125–144.

Gronau, R. (1973). Wage Comparisons: a Selectivity Bias. *NBER Working Paper №13.*

Heckman, J. (1974). Shadow Prices, Market Wages, and Labor Supply. *Econometrica: journal of the econometric society*, 679-694.

Heckman, J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Sample Estimator for Such Models. *Annals of Economic and Social Measurement*, 5(4), 475–492.

Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica: Journal of Econometric Society*, 47(1), 153–161.

Heckman, J. (1990). Varieties of Selection Bias. *The American Economic Review*, 313-318.

Heckman, J., Robb Jr, R. (1985). Alternative Methods for Evaluating the Impact of Interventions: An Overview. *Journal of Econometrics*, *30*(1), 239-267.

Khmelnitskaya, M. (2014). Russian housing finance policy: state-led institutional evolution. *Post-Communist Economies*, 26(2), 149–175.

LaCour-Little, M. (2007). The Home Purchase Mortgage Preferences of Low- and Moderate-Income Households. *Real Estate Economics*, 35, 265-290.

Munnell, A., G. Tootell, L. Browne, McEneaney, J. (1996). Mortgage Lending in Boston: Interpreting HMDA Data. *American Economic Review*, 86, 25–53.

Newey, W. K. (1997). Convergence Rates and Asymptotic Normality for Series Estimators. *Journal of Econometrics*, *79*(1), 147-168.

Newey, W. K. (1999). Two-step Series Estimation of Sample Selection Models. *Working paper*, MIT, Department of Economics.

Newey, W. K. (2013). Nonparametric Instrumental Variables Estimation. *The American Economic Review*, *103*(3), 550-556.

Newey, W. K., Powell, J. L. (1989). Nonparametric Instrumental Variables Estimation. *Working paper*, MIT, Department of Economics.

Newey, W. K., Powell, J. L., Vella, F. (1999). Nonparametric Estimation of Triangular Simultaneous Equations Models. *Econometrica*, *67*(3), 565-603.

Ozhegov E.M. (2014). Modelling Demand for Mortgage Loans Using Loan-Level Data. In: S.V. Ivliev, A.K. Bera, F.Lillo (eds.). *Financial Econometrics and Empirical Market Microstructure*, Springer.

Ozhegov E. M., Poroshina A. M. (2013). The Lagged Structure of Dynamic Demand Function for Mortgage Loans in Russia. *EJournal of Corporate Finance*, 27, 37-49.

Phillips, R., Yezer, A. (1996). Self-Selection and Tests for Bias and Risk in Mortgage Lending: Can You Price the Mortgage If You Don't Know the Process? *Journal of Real Estate Research*, 11, 87–102.

Rachlis, M., Yezer A. (1993). Serious Flaws in Statistical Tests for Discrimination in Mortgage Markets. *Journal of Housing Research*, 4, 315–336.

Ross, S.L. (2000). Mortgage Lending, Sample Selection and Default. *Real Estate Economics*, 8, 581–621.

Sanderson, E., Windmeijer, F. (2014). A Weak Instruments F-test in linear IV models with multiple endogenous variables. *Discussion paper 14/644*, University of Bristol, Department of Economics.

Stock, J.H., Yogo M. (2005). Testing for weak instruments in linear IV regression. In: D.W.K. Andrews and J.H. Stock (Eds.), *Identification and Inference for Econometric Models, Essays in Honor of Thomas Rothenberg*, 80-108. New York: Cambridge University Press.

Vella, F. (1998). Estimating Models with Sample Selection Bias: A Survey.*Journal of Human Resources, 33*(1).

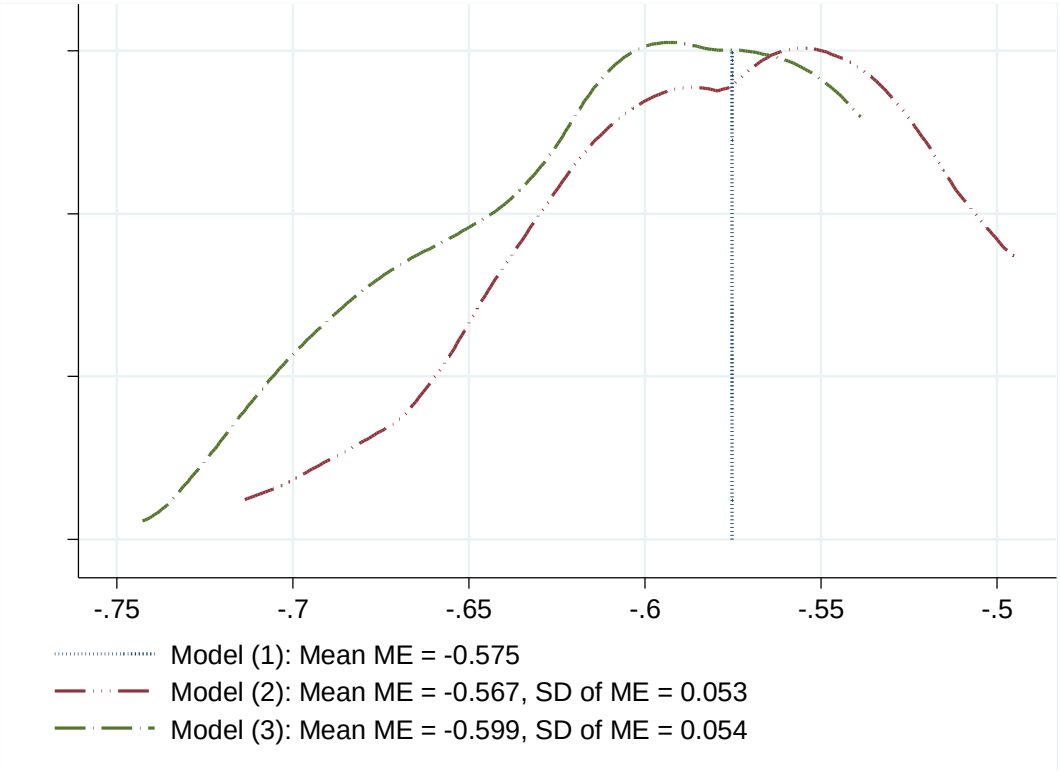
Yezer, A., Philips, R., Trost R. (1994). Bias in Estimates of Discrimination and Default in Mortgage Lending: the Effects of Simultaneity and Self-Selection. *Journal of Real Estate Finance and Economics*, 9, 197–215.

Appendix

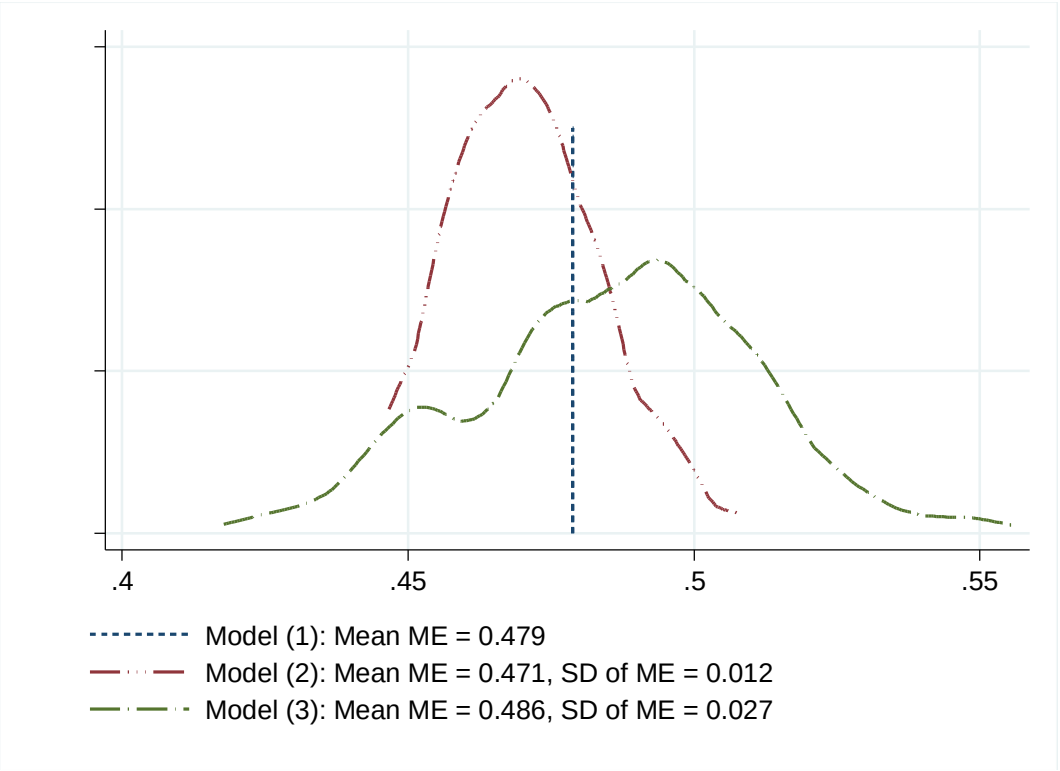
Tab. A1. Estimated parameters for selection equation.

Variable	(1) OLS	(2) Probit
Age of borrower	-0.006 (0.009)	-0.014 (0.025)
Age squared	0.000 (0.000)	0.000 (0.000)
Male	0.028 (0.018)	0.081 (0.051)
Family status (Married is base level):		
Single	-0.029 (0.025)	-0.093 (0.071)
Divorced	-0.042 (0.029)	-0.130 (0.083)
Widowed	-0.130* (0.076)	-0.363* (0.209)
Category of activity (Hired employee is base level):		
Entrepreneur	0.066 (0.099)	0.165 (0.294)
State employee	0.140*** (0.045)	0.393*** (0.131)
Level of education (Complete higher is base level):		
Less than higher	-0.071*** (0.017)	-0.197*** (0.047)
Number of co-borrowers (No co-borrowers is base level)		
1 co-borrower	0.001 (0.024)	-0.015 (0.069)
2 co-borrowers	0.019 (0.048)	0.055 (0.140)
Declared income of co-borrowers (Not declared is base level):		
From 0 to $249	0.155*** (0.052)	0.731*** (0.198)
From $250 to $499	0.088** (0.043)	0.291** (0.135)
More than $500	0.073 (0.045)	0.245* (0.138)
Declared income of main borrower (Not declared is base level):		
From 0 to $249	-0.011 (0.054)	-0.083 (0.151)
From $250 to $499	0.265*** (0.034)	0.798*** (0.107)
From $500 to $999	0.232*** (0.027)	0.656*** (0.080)
More than $1000	0.179*** (0.036)	0.475*** (0.105)
Difference between AHML loans number and number of applications	-0.000*** (0.000)	-0.001*** (0.000)
Constant	0.646*** (0.161)	0.295 (0.452)
N	3344	3344
k	20	20
% of correct predictions	64.8	64.7
Test for excluded variable significance	$F(1, 3224)$=31.98	$\chi^2(1)$=32.23

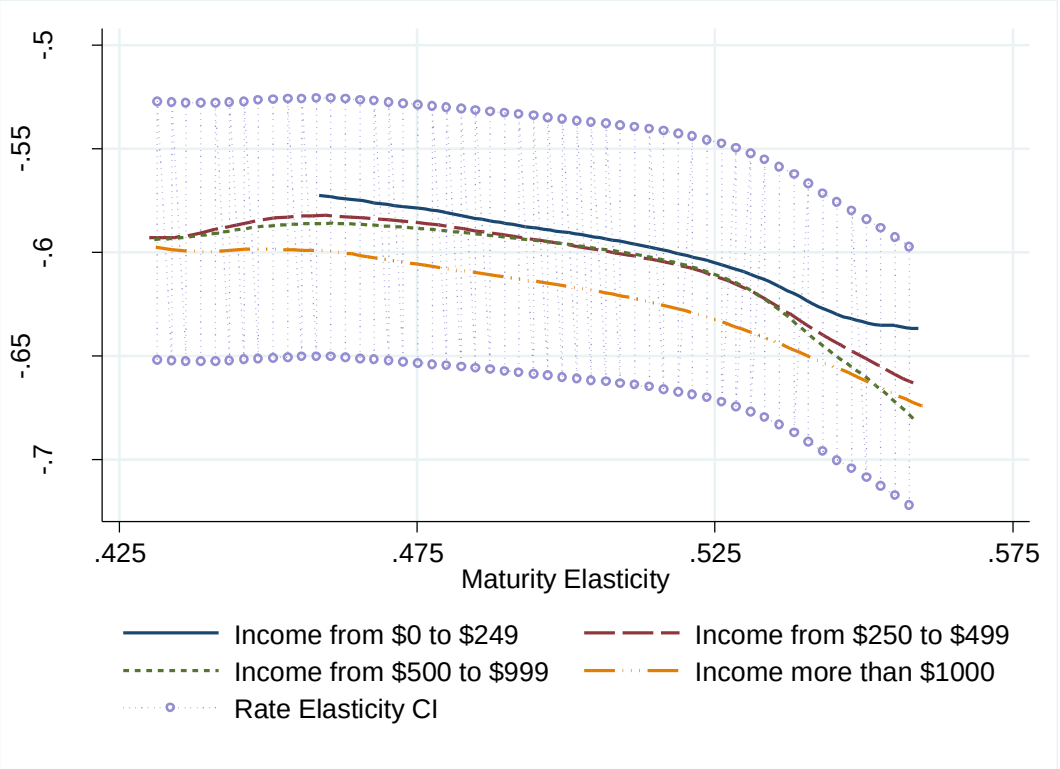
Note: Robust standard errors are in parenthesis, significance level obtained from *t*-statistics,
* - 10%, ** - 5%, *** - 1%.
k – number of estimated parameters, N – number of observations



Graph 1. Distribution of demand elasticity on interest rate



Graph 2. Distribution of demand elasticity on maturity



Graph 3. Demand elasticities on interest and maturity for different income groups

Lemma 1. *If functions* $g_0(x,z_0)$, $\gamma_j(x,z)$, $\lambda_j(p)$ *are continuously differentiable with continuous distribution functions almost everywhere and with probability one* $\frac{\partial g_0(x,z_0)}{\partial z_0} \neq 0$, *then* $\gamma_j(x,z)$ *is identified up to an additive constant.*

Proof (is similar to T.2.1 in Das et al. (2003)): Any observationally equivalent model for (3) must has $\left[y_j \middle| x,z,z_0,d=1\right] = \hat{\gamma}_j(x,z) + \hat{\lambda}_j(p)$. Consider $f_1(x,z) + f_2(p) = 0$, where $f_1(x,z) = \gamma_j(x,z) - \hat{\gamma}_j(x,z)$, and $f_2(p) = \lambda_j(p) - \hat{\lambda}_j(p)$. If g_0, γ_j and λ_j are differentiable, then f_1 and f_2 are also differentiable. Then we may differentiate $f_1 + f_2 = 0$ by the set of (z_0, x, z):

$$\begin{aligned} 0 &= \frac{\partial f_2(p)}{\partial p}\frac{\partial g_0(x,z_0)}{\partial z_0} \\ 0 &= \frac{\partial f_1(x,z)}{\partial x} + \frac{\partial f_2(p)}{\partial p}\frac{\partial g_0(x,z_0)}{\partial x} \\ 0 &= \frac{\partial f_1(x,z)}{\partial z} \end{aligned} \qquad \text{(A.1)}$$

First condition and $\frac{\partial g_0(x,z_0)}{\partial z_0} \neq 0$ imply $\frac{\partial f_2(p)}{\partial p} = 0$, then f_2 is constant.

Then the second condition gives $\frac{\partial f_1(x,z)}{\partial x} = 0$. It means that $f_1(x^1,z)$ is constant and $\hat{\pi}(x^1,z) = \pi(x^1,z) + C$. ||

Proof of Theorem 1. By lemma 1 equations (2-3) is identified. Let us prove the identification of equation (4). Any observationally equivalent model for (4) must has $E\left[y_j \middle| y_{-j},x,z,z_0,d=1\right] = \hat{g}_j\left(y_{-j},x,z_j\right) + \hat{\varphi}_j(p,e_{-j})$ Consider $f_5\left(y_{-j},x,z_j\right) + f_6(p,e_{-j}) = 0$, where $f_5\left(y_{-j},x,z_j\right) = g_j\left(y_{-j},x,z_j\right) - \hat{g}_j\left(y_{-j},x,z_j\right)$ and $f_6(p,e_{-j}) = \varphi_j(p,e_{-j}) - \hat{\varphi}_j(p,e_{-j})$.

If $g_0, \gamma_j, g_j, \lambda_j, \varphi_j$ are continuously differentiable then $f_5\left(y_{-j}, x, z_j\right)$ and $f_6(p, e_{-j})$ are also continuously differentiable then we may differentiate $f_5 + f_6 = 0$ by the set of exogenous variables $\left(z_j, z_{-j}, x, z_0\right)$:

$$\begin{aligned}
0 &= \frac{\partial f_5\left(y_{-j},x,z_j\right)}{\partial z_j} + \frac{\partial f_5\left(y_{-j},x,z_j\right)}{\partial y_{-j}} \frac{\partial \gamma_{-j}(x,z)}{\partial z_j} \\
0 &= \frac{\partial f_5\left(y_{-j},x,z_j\right)}{\partial y_{-j}} \frac{\partial \gamma_{-j}(x,z)}{\partial z_{-j}} \\
0 &= \frac{\partial f_5\left(y_{-j},x,z_j\right)}{\partial y_{-j}} \frac{\partial \gamma_{-j}(x,z)}{\partial x} + \frac{\partial f_5\left(y_{-j},x,z_j\right)}{\partial x} + \frac{\partial f_6(p,e_{-j})}{\partial p} \frac{\partial g_0(x,z_0)}{\partial x} \\
0 &= \frac{\partial f_6(p,e_{-j})}{\partial p} \frac{\partial g_0(x,z_0)}{\partial z_0}
\end{aligned} \tag{A.3}$$

The last condition and $\frac{\partial g_0(x,z_0)}{\partial z_0} \neq 0$ imply that $\frac{\partial f_6(p,\nu,e_{-j})}{\partial p} = 0$.

Rank condition gives that for every j there is z_j with $\frac{\partial \gamma_j(x,z)}{\partial z_j} \neq 0$ and $\frac{\partial \gamma_{-j}(x,z)}{\partial z_{-j}} \neq 0$ respectively make the second condition equivalent to $\frac{\partial f_5\left(y_{-j},x,z_j\right)}{\partial y_{-j}} = 0$.

Replacing $\frac{\partial f_5\left(y_{-j},x,z_j\right)}{\partial y_{-j}} = 0$ in the third condition and using $\frac{\partial f_6(p,e_{-j})}{\partial p} = 0$ we have $\frac{\partial f_5\left(y_{-j},x,z_j\right)}{\partial x} = 0$.

And $\frac{\partial f_5\left(y_{-j},x,z_j\right)}{\partial y_{-j}} = 0$ in the first condition gives $\frac{\partial f_5\left(y_{-j},x,z_j\right)}{\partial z_j} = 0$.

All the obtained results imply that $f_5\left(y_{-j}, x, z_j\right) = g_j\left(y_{-j}, x, z_j\right) - \hat{g}_j\left(y_{-j}, x, z_j\right)$ is constant, consequently $\hat{g}_j\left(y_{-j}, x, z_j\right) = g_j\left(y_{-j}, x, z_j\right) + \mathrm{C}_j'$. ‖

Proof of Theorem 2. Consider a procedure of model (1) identification. It will take 3 steps:

1. On the first step we estimate the propensity score $p = E[d|x_0, w_0]$ from the selection equation:

$$d_i = \begin{cases} 1, g_0(x_i, z_{0i}) + e_{0i} \geq 0 \\ 0, g_0(x_i, z_{0i}) + e_{0i} < 0 \end{cases} \tag{A.4}$$

For every marginal distribution f_{e_0}, $E[d|x, z_0] = E[d = 1|x, z_0] = \int_{-g_0(x,z_0)}^{\infty} f_{e_0}(s)ds = \gamma_0(x, z_0)$. γ_0 with arbitrary distribution of e_0 and functional form of g_0 will be a function with arbitrary functional form but will depend only on the known set of variables, x, z_0.

We may decompose γ_0 into the Taylor series in a neighborhood of each (x_i, z_{0i}). $p_i = E[d_i|x_i, z_{0i}]$ may be approximated by a polynom $Q_0 = Q^{\rho_0}(x_i, z_{0i})\alpha_0$, where $Q^{\rho_0}(x, z_0)$ is polynomial approximating series for $\gamma_0(x, z_0)$ with ρ_0 and α_0 is a vector of parameters with dimensiality $\kappa = \frac{(\rho_0+\chi_0)!}{\rho_0!\chi_0!}$, $\chi_0 = dim(x, z_0)$.

Estimate of α_0 may be obtained by OLS as

$$\hat{\alpha}_0 = [Q_0{}'Q_0]^{-1}Q_0{}'d \tag{A.5}$$

For all fixed ρ_0 we may prove the consistency of $\hat{\alpha}_0$.

$$\begin{aligned}
\operatorname*{plim}_{n\to\infty} \hat{\alpha}_0 &= \operatorname*{plim}_{n\to\infty} [Q_0{}'Q_0]^{-1}Q_0{}'d = \operatorname*{plim}_{n\to\infty} [Q_0{}'Q_0]^{-1}Q_0{}'(Q_0\alpha_0 + \eta_0) \\
&= \alpha_0 + \operatorname*{plim}_{n\to\infty} [Q_0{}'Q_0]^{-1}Q_0{}'\eta_0 = \alpha_0
\end{aligned} \tag{A.6}$$

with the exogeneity of (x, z_0).This is obvious that a convergence speed to true $\gamma_0(x, z_0)$ depends on the power ρ_0 of approximation function. The higher ρ_0 gives the slower speed of convergence due to increase in the number of parameters being estimated. Das et al. (2003) showed that with the upper limit to an approximation polynom power the estimate is asymptotically normal. In this paper we will not prove the asymptotic normality and point out that standard errors may be obtained by bootstrap. The basics of asymptotic theory for two-step correction procedures provided by Newey (1997). It is also mentioned in

Das et al. (2003) that regression function may be represented as partially linear in regressors then all identification conditions should be held only for nonlinear part of regression function. Then the assumption of differentiability of regression functions may be relaxed when we include all discrete regressors only to linear part of regression function.

Then the propensity score will be

$$\widehat{p_i} = E[d_i|x_i, z_{0i}] = Q_0[Q_0'Q_0]^{-1}Q_0'd_i \tag{A.7}$$

2. On the second step we estimate the reduced form residuals corrected for sample selection:

$$e_j = y_j - E\big[y_j\big|x, z, z_0, d = 1\big] \tag{A.8}$$

If e_j has joint marginal distribution with e_0 with density function f_{e_0,e_j} then

$$E[e_j|x, z, z_0, d = 1] = E[e_j|g_0(x, z_0)+e_0 \geq 0] = \int_{-\infty}^{\infty}\int_{-g_0(x,z_0)}^{\infty} e_j f_{e_0,e_j}(s, r)dsdr = \lambda_j(p) \tag{A.9}$$

y_j is decomposed into regression and control functions:

$$y_j = \gamma_j(x, z) + \lambda_j(p) + \eta_j \tag{A.10}$$

The error term in this equation η_j is independent on (x, z).

If $\hat{p}$ is a propensity score then on this stage it will be fixed. (x, z) and $\hat{p}$ are two sets of different variables if $\frac{\partial Q^{\rho_0}(x_i,z_{0i})\widehat{\alpha}_0}{\partial z_0} \neq 0$.

Every arbitrary functions $\gamma_j(x, z)$ and $\lambda_j(\hat{p})$ may be approximated by $Q^{\rho_1}(x, z)b_{1j}$ and $Q^{\rho_1}(\hat{p})b_{2j}$ respectively, where $Q^{\rho_1}(x, z)b_{1j}$ and $Q^{\rho_1}(\hat{p})b_{2j}$ are polynomial approximating series with a power ρ_1, Then y_j may be approximated by the following equation:

$$y_j = Q^{\rho_1}(x, z)b_{1j} + Q^{\rho_1}(\hat{p})b_{2j} + \eta_j \tag{A.11}$$

Equation (A.17) is identified up to an additive constant when conditions of Theorem 1 are satisfied. Polynomial approximations for γ_j and λ_j satisfy differentiability condition. And we also need $\frac{\partial Q^{\rho_0}(x_i,z_{0i})\widehat{\alpha}_0}{\partial z_0} \neq 0$.

Let $b_j = (b_{1j}, b_{2j})$ and $Q_r = (Q^{\rho_1}(x, z), Q^{\rho_1}(\hat{p}))$ then b_j may be obtained by OLS as

$$\hat{b}_j = [Q_r'Q_r]^{-1}Q_r'y_j \tag{A.12}$$

With some large enough ρ_1, $Q^{\rho_1}(x, z)\hat{b}_{1j}$ is an approximation for $\gamma_j(x, z)$. And $\hat{b}_j = (\hat{b}_{1j}, \hat{b}_{2j})$ are consistent with independency of η_j and (x, z) due to

$$\operatorname*{plim}_{n\to\infty}\hat{b} = \operatorname*{plim}_{n\to\infty}[Q_r'Q_r]^{-1}Q_r'y_j = = \operatorname*{plim}_{n\to\infty}[Q_r'Q_r]^{-1}Q_r'\big(Q_r b_j + \eta_j\big) = = b_j + \operatorname*{plim}_{n\to\infty}[Q_r'Q_r]^{-1}Q_r'\eta_j = b_j \tag{A.13}$$

Identification of an additive constant in this equation is an additional research question when its true value is a point of interest. Heckman (1990) provided examples when identification of constant is essential. Andrews and Schafgans (1998) discussed also the identification strategy. When the identification of constant is not a point of interest then we only need to fix a value of some parameter. For example, let the parameter behind $(\hat{p})^0$ in $Q^{\rho_1}(\hat{p})$ be equal to 0. On the next steps we will also put 0 as a value of parameter behind the polynomial term with 0 power in control function.

Then the reduced form residuals will be

$$\hat{e}_{ji} = y_{ji} - Q_r\hat{b}_j \tag{A.14}$$

3. On the third step we estimate the structural equations corrected for sample selection and simultaneity in *y*.

If e_j has joint distribution with e_0 and e_{-j} with density function $f_{e_0,e}$ then

$$E[e_j \left| y_{-j}, x, z, z_0, d = 1 \right.] = E[e_j \left| e_{-j}, g_0(w_0, x_0) + e_0 \geq 0 \right.] = \int_{-\infty}^{\infty} \int_{-g_0(w_0,x_0)}^{\infty} e_j f_{e_0,e}(s, r \left| e_{-j} \right.) ds dr = \varphi_j(p, e_{-j}) \quad \text{(A.15)}$$

y_j is decomposed into

$$y_j = g_j\left(y_{-j}, x, z_j\right) + \varphi_j(p, e_{-j}) + \varepsilon_j \quad \text{(A.16)}$$

The error term ε_j in this equation will be independent on $\left(y_{-j}, x, z_j\right)$.

If $\hat{p}$ is the propensity score and $\hat{e}_{-j}$ are reduced form residuals then $\hat{p}$ and $\hat{e}_{-j}$ on this step are fixed. And (y_{-j}, x, z_j) and $(\hat{p}, \hat{e}_{-j})$ are sets of different variables if $\frac{\partial Q^{\rho_0}(x_i, z_{0i})\hat{\alpha}_0}{\partial z_0} \neq 0$ and $rank\left[\frac{\partial Q^{\rho_1}(x,z)\hat{b}_1}{\partial z}\right] = dim(y)$.

Every functions $g_j\left(y_{-j}, x, z_j\right)$ and $\varphi_j(p, e_{-j})$ may be approximated by $Q^{\rho_1}\left(y_{-j}, x, z_j\right)\beta_{1j}$ and $Q^{\rho_1}(\hat{p}, \hat{e}_{-j})\beta_{2j}$ respectively, where $Q^{\rho_1}\left(y_{-j}, x, z_j\right)$ and $Q^{\rho_1}(\hat{p}, \hat{e}_{-j})$ are polynomial approximating series with a power ρ_1. Then y_j may be approximated by

$$y_j = Q^{\rho_1}\left(y_{-j}, x, z_j\right)\beta_{1j} + Q^{\rho_1}(\hat{p}, \hat{e}_{-j})\beta_{2j} + \varepsilon_j \quad \text{(A.17)}$$

Equation (A.23) is be identified up to an additive constant if Theorem 1 conditions are satisfied. Polynomial approximations for g_j and φ_j satisfy differentiability condition. And we also need $\frac{\partial Q^{\rho_0}(x_i, z_{0i})\hat{\alpha}_0}{\partial z_0} \neq 0$ and $rank\left[\frac{\partial Q^{\rho_1}(x,z)\hat{b}_1}{\partial z}\right] = dim(y)$.

Let $\beta_j = (\beta_{1j}, \beta_{2j})$ and $Q_j = \left(Q^{\rho_1}\left(y_{-j}, x, z_j\right), Q^{\rho_1}(\hat{p}, \hat{e}_{-j})\right)$ then the estimate for β_j may be obtained by OLS as

$$\hat{\beta}_j = [Q_j{}' Q_j]^{-1} Q_j{}' y_j \quad \text{(A.18)}$$

For some large enough ρ_1, $Q^{\rho_1}\left(y_{-j}, x, z_j\right)\hat{\beta}_{1j}$ will be an approximation for $g_j\left(y_{-j}, x, z_j\right)$. Estimate $\hat{\beta}_j = (\hat{\beta}_{1j}, \hat{\beta}_{2j})$ is consistent with independence of ε_j and $\left(y_{-j}, x, z_j\right)$ due to

$$\operatorname*{plim}_{n\to\infty} \hat{\beta}_j = \operatorname*{plim}_{n\to\infty} [Q_j{}' Q_j]^{-1} Q_j{}' y_j = \operatorname*{plim}_{n\to\infty} [Q_j{}' Q_j]^{-1} Q_j{}' (Q_j \beta_j + \varepsilon_j) = \beta_j + \operatorname*{plim}_{n\to\infty} [Q_j{}' Q_j]^{-1} Q_j{}' \varepsilon_j = \beta_j \| \quad \text{(A.19)}$$